Pol. J. Environ. Stud. Vol. 30, No. 4 (2021), 1-10
DOI: *10.15244/pjoes/130677* ONLINE PUBLICATION DATE:

Original Research

Spatial Torrential Rainfall Modelling in Pattern Analysis Based on Robust PCA Approach

Shazlyn Milleana Shaharudin[1]*, Siti Mariana Che Mat Nor[1], Mou Leong Tan[2], Mohd Saiful Samsudin[3], Azman Azid[4], Shuhaida Ismail[5]

[1]Universiti Pendidikan Sultan Idris, Department of Mathematics, Faculty of Science and Mathematics, Malaysia
[2]Geoinformatic Unit, Geography Section, School of Humanities, Universiti Sains Malaysia, 11800 USM, Pulau Pinang, Malaysia
[3]Faculty Business and Entrepreneuship, Universiti Malaysia Kelantan, Kampus Kota, Karung Berkunci 36, Pangkalan Chepa, 16100 Kota Bharu, Kelantan, Malaysia
[4]Faculty of Bioresources and Food Industry, Universiti Sultan Zainal Abidin, Besut Campus, 22200 Besut, Terengganu, Malaysia
[5]Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Malaysia

Received: 20 August 2020
Accepted: 19 November 2020

Abstract

In this research work, the pattern of spatial cluster had been identified for torrential rainfall data within the context of Peninsular Malaysia, which experiences heavy pour annually. Hence, a robust Principal Component Analysis (PCA) technique was employed in this study in order to address problem related to non-balance cluster(s) across patterns of rainfall stemming from skewed rainfall data. To analyze the observations made, Tukey's biweight correlation was applied. For PCA components extraction, the optimum breakdown point was determined based on the proposed method. In order to strike a balance for extraction of number of components, as well as to hinder insignificant spatial scale or low-frequency variation, the simulation data recorded a breakdown point at 70% cumulative percentage of variance. The study outcomes revealed that the robust PCA gave better enhancement than the Pearson-based PCA did for cluster average number and quality. The findings indicate that ten rainfall patterns obtained are quite definite and clearly display the dominant role extended by the complex topography and exchange monsoons of the peninsular.

Keywords: Principal Component Analysis (PCA), robust PCA, Tukey's biweight correlation, Pearson Correlation, K-means Cluster Analysis

*e-mail: shazlyn@fsmt.upsi.edu.my

Introduction

Continuous torrential rainfall in Malaysia may pose as a calamity threat, such as the worst flood event that hit Kelantan on 28th December 2014 [1]. As a result, Malaysian meteorologists assess patterns of rainfall by emphasizing on heavy pour. The outcomes serve as a guide for devising effective actions and viable precautions to prevent flood.

Studies pertaining to spatial rainfall patterns within the hydrology domain have typically applied two techniques; regression and clustering-based modelling. Some research work that has employed the regression approach [2-6] had characterized the patterns of rainfall distribution. This approach particularly detects trends than describes the regional attributes of the rainfall pattern. The outcomes of this regression approach can be used for forecasting purposes. On the other hand, the clustering approach determines both temporal and spatial rainfall patterns to describe the attributes of regional data, thus implying highly-structured rainfall patterns [7]. This method is a viable statistical tool for grouping of regions, as well as to identify periods of rainfall events in regions.

The standard clustering approach to detect patterns of spatial rainfall is inapt for tropical climate that receives plenty of rainfall in a year. The extended observation period accumulates a massive dataset with intricate, redundant, and irrelevant data, thus giving inaccurate results. Despite the wet and dry seasons, rainfall in tropical regions does not vary significantly when compared to regions with four seasons. Besides, noise present in huge rainfall datasets is bound to cause errors. Hence, identifying cluster pattern from a huge rainfall dataset is indeed challenging.

A technique of robust PCA is proposed to address the above mentioned problem. First, Tukey's biweight correlation was implemented in robust PCA for measuring scale and location can down-weigh observations distant from data centre and has resistance against outlying observations due to the features of the tool [8-9]. Second, as breakdown point is integral for identifying the optimum PCA components to be extracted, a new breakdown set is prescribed for comparison with the amount of extracted components so as to strike a balance for extraction of integral components.

Pearson-based PCA performance was compared with that of the proposed PCA based on Tukey's biweight using simulated dataset matrices that reflected the real rainfall dataset. This comparison identified the spatial cluster pattern for heavy rainfall recorded in Peninsular Malaysia.

Materials and Methods

Study Area

Rainfall data (measured using bucket rain gauge) recorded in 75 stations across Peninsular Malaysia were retrieved from Jabatan Pengairan dan Saliran (JPS) as

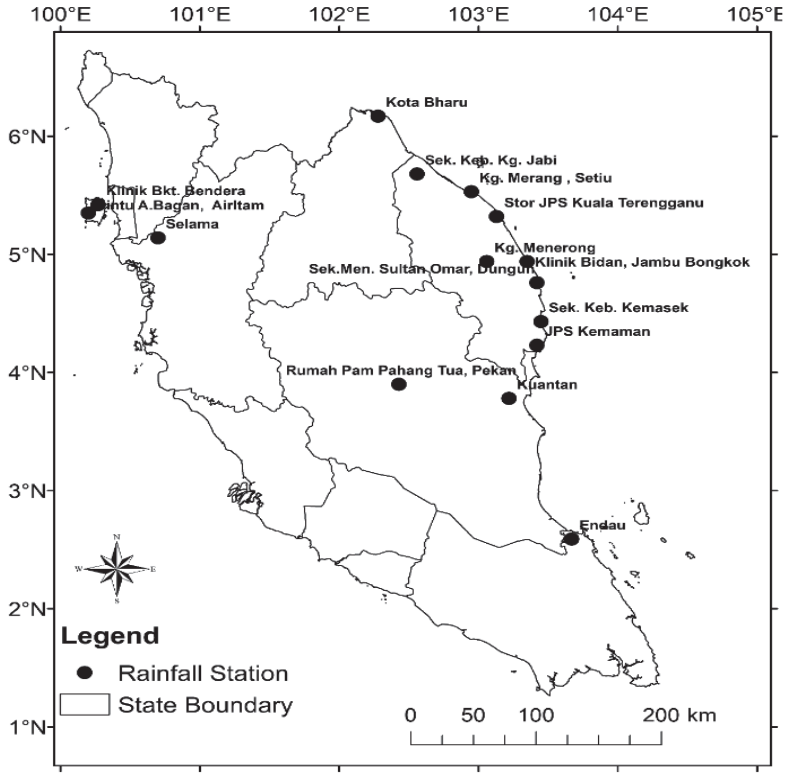

Fig 1. Rainfall stations across the main torrential centres located in Peninsular Malaysia.

shown in Fig. 1. This dataset, which is without missing data and consisted of 903,375 daily measurements from 1975 to 2007 (12,045 days, excluding 8 leap days), had been deemed sufficient to detect patterns of torrential rainfall.

Since the focus of this present study is on torrential rainfall, a set of criteria was prepared to establish a threshold and to distinguish torrential rainfall across selected regions from the rest. As a result, it was decided that 60 mm/day as the threshold for torrential rainfall for the context of this study [10]. Filtering the data using the criteria for 2% of the total stations yielded 15 stations and 250 days, which had been adequately dense for determining spatial dissemination across regional scale [11].

Principal Component Analysis

PCA refers to a commonly used statistical measuring instrument for reduction of data across vast domains, including hydrology, image compression, and climatology. The PCA is typically applied to minimize the amount of variables into smaller component groups while concurrently retaining essential data. Here, principal components (linearly uncorrelated variables) are yielded from conversion of observed probable correlated variables. According to [12-13], this particular technique has been reckoned as effective to detect data with high dimensionality.

The original data variation is essential in identifying the initial principal component. The subsequent component(s) reflect the remainder uncorrelated variation with prior component(s). Stemming from data matrix, correlation matrix (covariance) has a vital function in PCA to determine both eigenvectors and eigenvalues in identifying components that are related to represent data variations [10]. The following presents the correlation coefficient of Pearson for two observed vectors:

$$r_{ij} = \frac{\sum_{i=1}^{n}(X_i-\bar{X}_i)(X_j-\bar{X}_j)}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X}_i)^2 \sum_{i=1}^{n}(X_j-\bar{X}_j)^2}} \quad (1)$$

...where $\bar{X}_i$ and $\bar{X}_j$ denote vector mean values, while X_i and X_j are observed vectors in matrix data, X, along with n observations. According to [10], the threshold of eigenvalues for a huge dataset to extract components is 70% of total variance. Minimized matrix refers to the eigenvector loading component matrix that reflects a new set of variables with original variables being transformed linearly, which maximizes variance at new axes.

The Pearson-based PCA has been commonly applied for Eigen analysis in order to yield components mostly linked with data variation. Nevertheless, this method has been proven sensitive towards data without Gaussian distribution, along with observations that are skewed (e.g., outlying values). According to [14-15], climate-related data, particularly rainfall data, have distributions that are skewed towards positive values that are high. As Pearson approach provides weights that are equal for dataset observation, this technique is not robust towards outliers. Therefore, this Pearson-based PCA method can influence cluster partitions and cause cluster imbalance within high-dimensional space.

In order to address the above mentioned issue, Tukey's biweight robust correlation was employed for identifying rainfall pattern in Peninsular Malaysia. This is due to the fact that the proposed technique has better resistance against outlying values by evaluating every observation and down-weighing those distant from data centre.

PCA Based Tukey's Biweight Correlation

Tukey's biweight correlation was initially proposed by [16] to analyze microarray gene expression data. In general, the method consists of a basic algorithm and several variation of the algorithm to further fine-tune its correlation measure. This correlation measure is based on Tukey's biweight and can be used both in clustering and gene network algorithm. The entries in the data of the applied method represent the sums of the contributions in measures of similarity in correlation approach. The Turkey's biweight is one of the family of M-estimators are used to estimate location and scatter. This approach works iteratively using a weight function that makes it more resistant to outlying values, where it down weights those that lies far from the center of the data. Another important part in Tukey's biweight correlation is a breakdown point. According to the study, the breakdown point is used in measuring their resistance to outlying data values [17]. However, in PCA based Tukey's biweight correlation, the breakdown point is used to determine the best number of components to extract.

Upon relying on M-estimators, Tukey's biweight correlation predicts robust correlation. These M-estimators possess a function that is derivative in identifying assigned weights to the dataset. Hence, observations can be down-weighed in order to portray data centre impact [15]. This function of derivative is expressed in the following:

$$\psi(u)\begin{cases} u(1-u)^2 & |u| \leq 1 \\ 0 & |u| > 1 \end{cases} \quad (2)$$

...where $\psi(u)$ is a biweight function and u represents the transformation of observations.

Upon $|u|$ being sufficiently massive, $\psi(u)$ is reduced to a value of zero. Breakdown point is essential to be measured in determining resistance towards outlying M-estimator data values. Breakdown point refers to the smallest contamination fraction that could yield inaccurate outcome [16]. Upon comparing several breakdown points (0.0, 0.2, 0.4, & 0.5) with Tukey's biweight, breakdown point 0.4 emerged as

the best as it resulted in more efficient and accurate yields [18].

The biweight correlation was yielded after determining location estimate, $\widetilde{T}$, and later, shape estimate, $\widetilde{S}$. The (i,j)th element of $\widetilde{S}$, such as $\tilde{s}_{ij}$, denotes the covariance resistant estimate for dual vectors; X_i and X_j. The following determines biweight correlation for the dual vectors:

$$\tilde{r}_{ij} = \frac{\tilde{s}_{ij}}{\sqrt{\tilde{s}_{ii}\tilde{s}_{jj}}} \tag{3}$$

and

$$T_n^{(k+1)} = \frac{\sum_{i=1}^{n} X_i w\left(u_{i(k)}\right)}{\sum_{i=1}^{n} w\left(u_{i(k)}\right)} \quad k = 0,1,2,\ldots \tag{4}$$

$$S_n^{(k+1)} = \frac{\sum_{i=1}^{n} w\left(u_{i(k)}\right)\left(X_i - T^{(k+1)}\right)\left(X_i - T^{(k+1)}\right)^t}{\sum_{i=1}^{n} w\left(u_{i(k)}\right)\left(u_{i(k)}\right)} \tag{5}$$

...where $T_n^{(k+1)}$ refers to vector location, whereas $S_n^{(k+1)}$ represents shape matrix, for instance k = 0,1,2,... In K-means cluster analysis, the proposed method enhanced partition of cluster, which had better resistance against values that are outlaid, in comparison to conventional PCA (Person correlation).

Breakdown point is one of the important aspects in M-estimator for resistance to outlying data values. There are two types of breakdown points: replacement breakdown and additive breakdown. The replacement breakdown executes when at least one of the original data points have been replaced with an arbitrary value to determine the performance of the estimator. The additive breakdown would show the performance of the estimator when random data are added to the original data set. In this study, we are concerned with the replacement breakdown which is the smallest fraction of a data set that one could replace with outlier values in the data set under any circumstances to take the estimator over all bounds [19]. The breakdown point for Tukey's biweight correlation can be adjusted over a range of values where it is not identical as the breakdown point for sample mean which is 0 and the breakdown point for median which is close to 1/2. Adjustments of the breakdown point will have effect in determining the number of components to obtain in PCA method. Thus, in this study, we tested the performance of the biweight correlation under a variety of different point change in order to determine the best number of components to extract in PCA to identify torrential rainfall pattern. Fig. 2. depicts the nine steps of proposed approached was applied to daily torrential rainfall data set in identifying rainfall patterns in Peninsular Malaysia.

Simulation of Robust PCA

In determining the effectiveness of the proposed PCA, a comparison was made between conventional and proposed PCA approaches using simulated data matrices that reflected real multivariate data of heavy rainfall within the context of Peninsular Malaysia. Data distribution for tropical rainfall were skewed towards right, wherein heavy rainfall modeling may employed these features. The multivariate rainfall data were tested using three techniques, namely Generalized Pareto distribution (GPD), log-normal, and gamma,

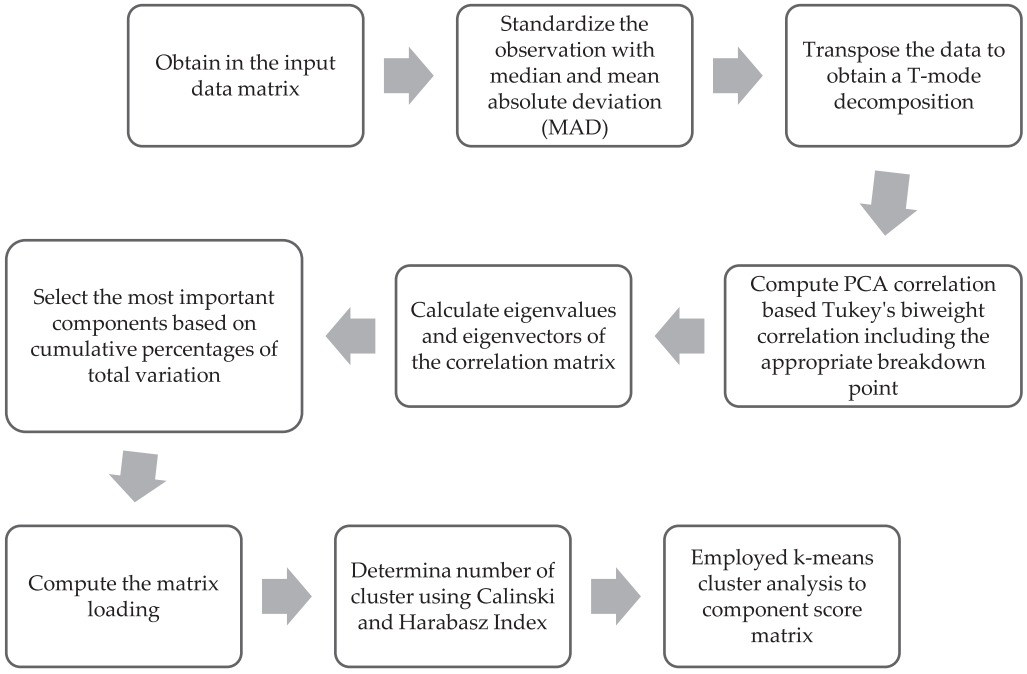

Fig 2. Nine steps of proposed approach was applied on daily torrential rainfall in Peninsular Malaysia.

which have been vastly employed for analyses of rainfall data [20-22]. Estimation of parameters for the selected probability distributions had been based on statistical summary of real data. As a result, GPD had best fit the dataset after a number of distribution graph evaluations and goodness of fit assessments involving Anderson Darling and Chi-square tests. As GPD was exceptional at significance level, the null hypothesis is not rejected (GPD gives accurate statistical model).

The simulations had been performed upon samples with distributions of GDP using parameters; shape ($\xi = 0.2$), location ($\mu = 104.8$), and scale ($\sigma = 54.7$), which were retrieved from the 33-year real heavy rainfall data. This led to the development of $n \times p$ matrix to denote 15 rainfall stations and 250 torrential rainfall days. Two settings were applied to vary the simulation test. First, after assessing the data variations, 60 mm/day was selected as the threshold in this study. Second, a breakdown point of 0.4 was selected for extraction of significant components in PCA.

Every generated dataset was tested using Pearson-based PCA and the proposed PCA. The outcomes retrieved from both PCA methods were compared in terms of cluster partition generation of imbalanced cluster rainfall patterns. Fig. 3. illustrated clearly the simulation procedures of proposed PCA approach in this study.

Evaluating Performance of Robust PCA

This study is particularly interested in using the results on number of cluster pattern to obtain and determine the appropriate breakdown point to evaluate the performance of the proposed Robust PCA based Tukey's biweght correlation. In order to reduce the random effect on the results, 20 simulated data were performed on each of the different settings described in Section 3.4.

As a result, for each data set $m_k^{(1)},...., m_k^{(20)}$ number of components were obtained to be extracted based on 70% cumulative percentage at different range of breakdown points as well the number of clusters based on k-means. Then, the average number of cluster and components produced by the simulated data are compared. Performances displayed by the two approaches were examined by looking into clustering quality yielded from validity indices, namely Silhouette, Davies-Bouldin, and Rand Indices. Sufficient component and varied number of clusters obtained are generally favored. This is because such results are more interesting from hydrologists' point of view in identifying different clusters of rainfall patterns in characterizing different climate phenomenon.

In addition, the possible extent of the proposed method was investigated in practical applications on torrential rainfall data in Peninsular Malaysia. The effect of the number of components were observed and clusters produced by the recommended breakdown point from the simulated results on robust PCA based Tukey's biweight correlation. Fig. 4. depicted the process of evaluation performance of Robust PCA.

Results and Discussion

Performances between the conventional and proposed PCA approaches were compared by employing simulated data. As tabulated in Table 1, average components were retrieved by employing the proposed PCA using 20 simulated data. As a result, the

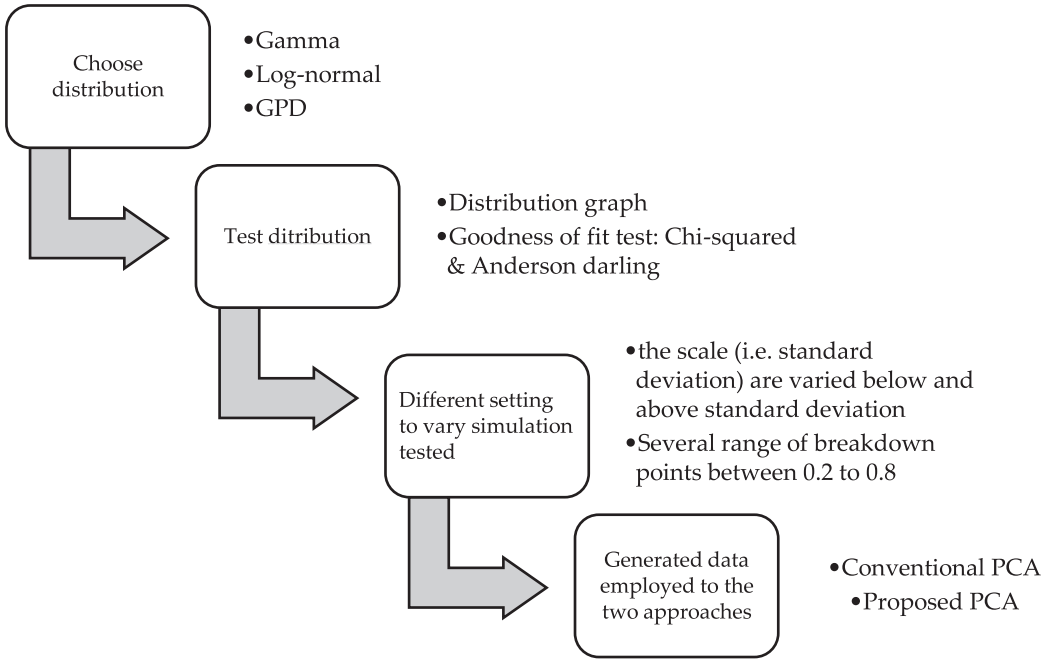

Fig 3. Simulation procedures of Robust PCA based on Tukey's biweight correlation.

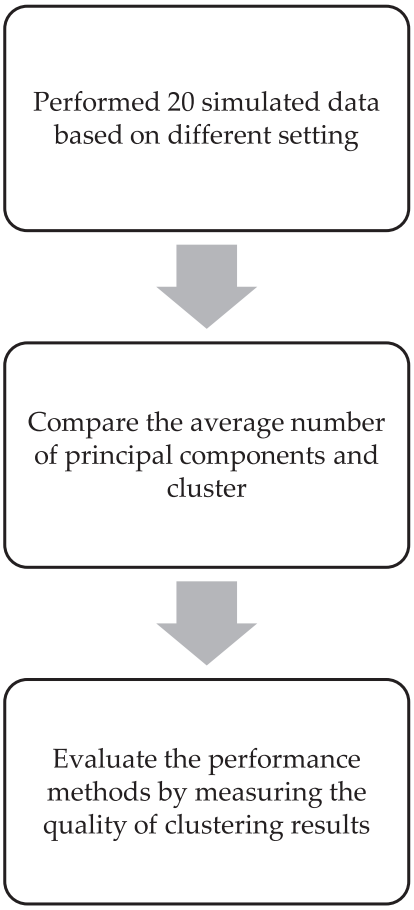

Fig 4. Evaluation performance of PCA based on Tukey's biweight correlation.

breakdown point had an impact on component number extraction for this proposed method. Table 1 shows that breakdown point with higher value (r = 0.8) extracted fewer essential components, while a breakdown point of 0.4 led to a good balance for extraction of 12 significant components. It is noteworthy to highlight that extracting plenty of components from hydrological data is undesired as it portrays insignificant spatial scale or low-frequency variations [23]. This emphasizes the importance of selecting the most viable breakdown point for the proposed PCA.

Tables 2 and 3 tabulate the average clusters and components retrieved from both conventional and proposed PCA upon increasing 60% to 80% of the cumulative percentage of variation. Both components and clusters (up to two decimal points) were retrieved from 20 simulated data. Variations displayed by the simulated data for every cluster, components, and cumulative percentage of variation had been small and ranged at 0.44-0.94.

Table 1. Average number of components based on 70% cumulative percentage of variance based on breakdown point values.

Breakdown Point, r	Number of Components
0.2	9
0.4	12
0.6	6
0.8	3

Table 2. Average number of components yielded by both PCA methods using 20 simulated data.

Cumulative percentage (%)	Number of components			
	Tukey's biweight		Pearson	
	Mean	Standard deviation	Mean	Standard deviation
60	2.25	0.44	45.40	0.82
65	5.55	0.76	54.05	0.89
70	11.55	0.94	61.50	0.83
75	19.80	0.89	71.55	0.89
80	28.75	0.92	82.50	0.69

Table 3. Average number of clusters yielded by both PCA methods using 20 simulated data.

Cumulative percentage (%)	Number of cluster, K			
	Tukey's biweight		Pearson	
	Mean	Standard deviation	Mean	Standard deviation
60	9.50	0.69	2.40	0.60
65	5.10	0.85	2.40	0.50
70	8.40	0.88	2.35	0.49
75	11.50	0.94	2.25	0.55
80	2.40	0.50	2.35	0.59

The differences between the two PCA approaches for the average number of components at every cumulative percentage of variations level are tabulated in Table 2. It appears that proposed PCA (Tukey's biweight correlation) requires less number of components to extract in order to achieve at least 70% of cumulative percentage of variation compared to conventional PCA. In an instance, 11.55≈12 components were extracted by the proposed PCA approach, while 61.50≈62 components for the conventional PCA at 70% cumulative percentage of variation. The presence of plenty of components when determining rainfall patterns is undesired as it suggests excessive outliers.

As for partition of clusters, the proposed PCA appeared to be more sensitive to cluster number in light of retained components, in comparison to the conventional PCA (see Table 3). The conventional PCA attained stability at two clusters at any cumulative

percentage of variation. Two clusters clearly is inappropriate as it mask the true structure of the data [24]. Nonetheless, it is sensible to identify more than two cluster partitions for detection and description for pattern of rainfall. Thus, conventional PCA is unsuitable to use in identifying rainfall patterns especially those countries which have similar climates to that of Malaysia.

In the next analysis, both PCA approaches were assessed using real heavy rainfall dataset. Findings presented in Table 4 were similar to those tabulated in Tables 2 and 3 for simulated data based on cumulative percentage of variance retained and the number of clusters obtained. Based on Table 4, the proposed PCA method required fewer components than the conventional PCA had needed for varying cumulative percentages of variation. The conventional PCA, as given in Table 4, only generated two clusters at any cumulative percentage of variation. The findings signify certain influential observations from the data. Nevertheless, the proposed PCA displayed a range of patterns for the cluster number generated at varying cumulative percentages of variation. This is ascribed to the clustering outcomes sensitivity towards component number yielded, wherein the number of components to be retained must be determined accordingly. According to [25], variation among clusters should indicate the path of a principal component, at least.

Serving as guideline, the higher values displayed by Rand and Silhouette indices, whereas the lower value by Davies-Bouldin index, reflect exceptional cluster quality. Table 5 presents the indices of average validity that determined clustering outcomes quality based on the 20 simulated data retrieved from both PCA approaches. The simulated data variation for every validity index had been small and ranged at 0.03-0.44. The proposed PCA resulted in better clustering outcomes for the three indices, in comparison to the Pearson-based PCA. This portrays that the proposed PCA is indeed a viably robust technique that can be applied for hydrology studies, particularly for torrential rainfall data analysis. The proposed PCA method exhibited substantial enhancement for partition of cluster, when compared to Pearson-based PCA, so as to hinder identifying imbalance and inaccurate clusters in rainfall analysis.

Table 4. Number of components retained in PCA and number of clusters yielded based on two PCA methods using real dataset.

Cumulative percentage (%)	Number of components		Number of cluster, *K*	
	Tukey's biweight	Pearson	Tukey's biweight	Pearson
60	11	12	12	2
65	13	14	12	2
70	15	19	10	2
75	22	26	6	2
80	28	35	2	2

Table 5. Indices that determined the clustering outcomes quality based on 20 simulated data.

Correlation	Rand Index		Silhouette Index		Davies-Bouldin Index	
	Mean	SD	Mean	SD	Mean	SD
Tukey's biweight	0.67	0.12	0.30	0.06	3.40	0.44
Pearson	0.43	0.06	0.06	0.03	5.78	0.30

*SD denotes standard deviation.

The main features of the clustering result are discussed to verify the distinction between the clusters with respect to their significant locations and period of monsoon occurrence for the torrential rainfall patterns (RP) based on the recommended settings in the previous outcome. In defining the spatial characteristics of torrential rainfall pattern in Peninsular Malaysia, ten clusters are obtained. Fig. 5a) until Fig. 5j) show that the daily rainfall composites for the RP 1 until RP 10 obtained in the classification of torrential events in the Northern and Eastern region in Peninsular Malaysia. These clusters are mapped out using ArcGIS software. Note that the torrential rainfall maxima locations could be clearly identified in a dark scale from the maps.

Fig. 5a) and Fig. 5b) illustrated that RP 1 and RP 2 exhibits moderate rainfall in the whole region in Peninsular Malaysia, with a general increase for most of the highland uplands, such as Bukit Bendera (Pulau Pinang). The maximum torrential rainfall for RP 1 occurs in the Pintu Air Bagan Air Itam (Pulau Pinang) and RP 2 in Bukit Bendera (Pulau Pinang), but the main feature of this particular pattern is the wide distribution of the torrential rainfall areas, which comprise the entire region.The majority of the days associated with these patterns occurred between November and March during the northeast monsoon with 63.5% (RP 1) and 68.2% (RP 2) of torrential rainfall. Consequently, the southwest monsoon recorded 20.8% (RP 1) and 17.6% (RP 2) of torrential rainfall between May and September. With 15.7% (RP1) and 14.1% (RP 2), the lowest case was the inter monsoon, occurring in April and October.

Fig. 5c) shows that RP 3 represents heavy rainfall over the east region with maximum torrential rainfall in Kota Bahru (Kelantan). In this pattern, the intensity of the torrential rainfall decreases from east to the southwest region. Torrential rainfall occurs mostly in Kelantan due to strong influence by the northeast monsoon and occurrence of sea breeze. Furthermore, Kota Bahru (Kelantan) is located near the coast. Therefore, this exposes the region to more rainfall during that period. The highest percentage of days

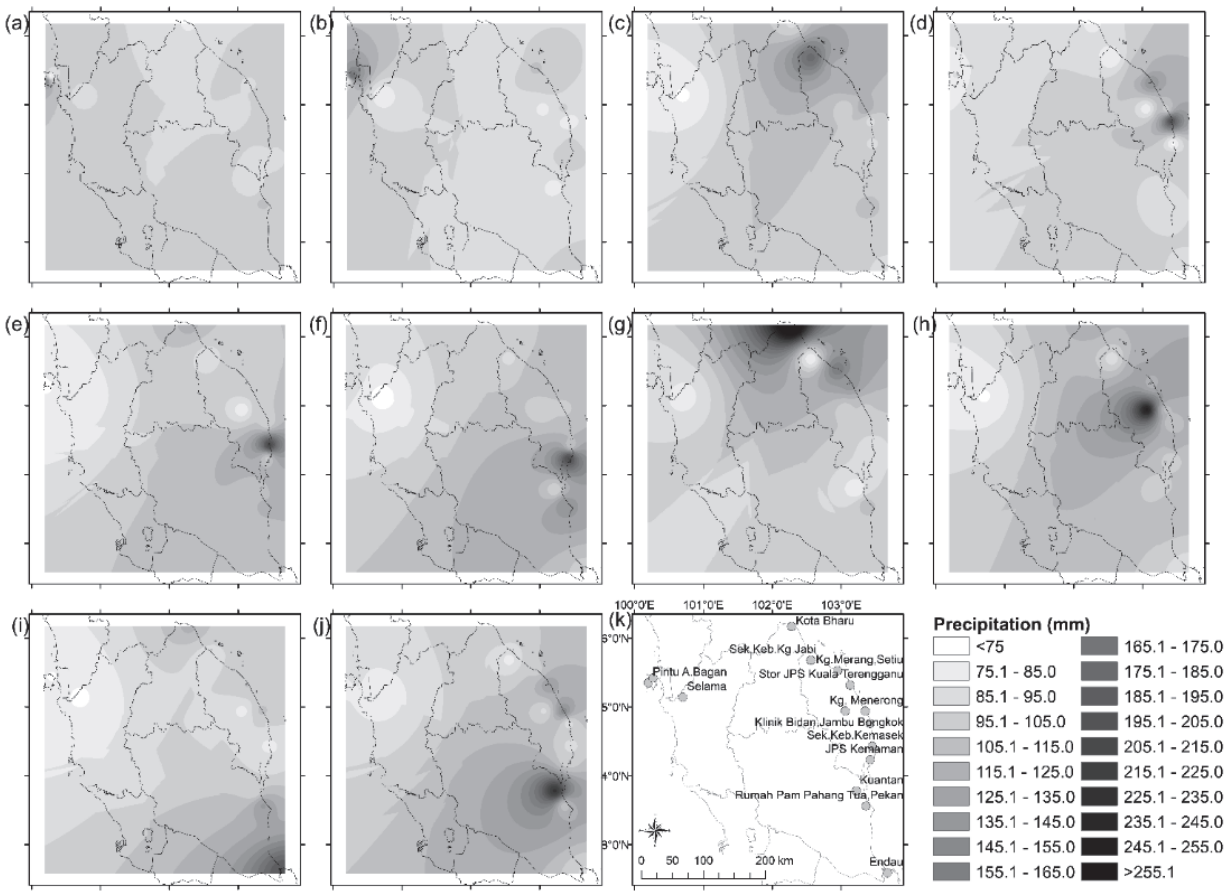

Fig. 5 Daily rainfall composites for the RP 1 until RP 10 obtained in the classification of torrential events in the Northern and Eastern region in Peninsular Malaysia.

associated with this pattern corresponds to the northeast monsoon with 66.7% of torrential days over 33 years, 9.4% of torrential rainfall in the inter monsoon seasons and 23.9% of torrential rainfall days occurred in the southwest monsoon season.

RP 4, RP 5, RP 6, RP 7 and RP 8 in Fig. 5d) until Fig. 5h) are characterized by heavy torrential and intense rainfall occurring in the Eastern region. All the patterns are located in Terengganu at different areas where RP 4 shows that the region receiving higher rainfall is in Dungun, RP 5 indicates that heavy rainfall occurs in Kemasek, RP 6 identifies Kemaman as the maximum torrential rainfall area, RP 7 identifies Kampung Jabi as the area with maximum torrential rainfall, and heavy torrential rainfall is shown to occur in Kampung Menerong in RP 8. Distribution of rainfall patterns for each group is significantly different due to different altitudes where the rainfalls are observed. Northwestern region received less rainfall because Titiwangsa Range blocks the moisture bearing clouds, that which possibly affects most of the rainfall stations along the western part of Peninsular Malaysia. Meanwhile, the Eastern part in Peninsular Malaysia is considered as wettest area due to the strong influenced by Northeast monsoon that bring heavy rainfall to the region in the period of November until March. The days included in all rainfall patterns in this study fall mostly in Northeast monsoon with almost the same percentage of the total cases of torrential rainfall over 33 years, 67%.

Fig. 5i) represents substantial rainfalls with maximums in Endau, Mersing where the topography is defined as lowland area. Naturally, water of the rainfall will flow from high to low area. Hence, when the northeast monsoon brings heavy rainfall to that area, this makes the location receive heavy rainfall. Furthermore, without ranges or mountains, the region is more likely to encounter rainfall. Due to the location concentrated close to the coast, the occurrence of sea breeze is also one of the major factors that cause this region to receive maximum rainfall. It can be seen clearly from Fig. 5i) , the pattern exhibits a gradual decrease from eastern to northern region. As in the other groups, the days included in this cluster consisted mostly of northeast monsoon with 62.1% of the total number of torrential rainfall cases. Meanwhile, Fig. 5j) shows that torrential rainfall of Kuantan (Pahang) is concentrated to urban area where it is characterized by higher population density and vast human features in comparison to areas surrounding it.

Table 6. Summary of the ten rainfall pattern groups obtained for daily torrential rainfall.

Rainfall Pattern	Region	Location	Days Included
RP 1	Northern	Pintu Air Bagan, Air Itam	17
RP 2		Bukit Bendera	17
RP 3	Eastern	Kota Bahru	18
RP 4		Dungun	19
RP 5		Kemasek	27
RP 6		Kemaman	29
RP 7		Kampung Jabi	41
RP 8		Kampung Menerong	32
RP 9		Endau	28
RP 10		Kuantan	22
			250

Normally, plants especially largest trees in urban area are difficult to find as urban areas undergo continuous built up of urban development that is within a labor market. Therefore, when northeast monsoon brings heavy rainfall in that region, it will directly receive heavy rainfall without any restriction from the largest trees. The region recorded higher torrential rainfall with 66.1% of these events occurring in the period of November until March.

As seen in Table 6, RP 7 pattern is significantly more frequent then the remainder RPs. This followed by Kampung Menerong with RP 8. The least RPs is RP 1 and RP 2 that received less torrential rainfall and both of these patterns are located in northern region. From Table 7, it can be seen clearly that northeast

Table 7. Percentage frequency distribution of torrential rainfall days over 33 years according to monsoon occurred for ten rainfall patterns.

Rainfall Pattern (RP)	SW Monsoon (%)	INTER Monsoon (%)	NE Monsoon (%)
RP 1	20.8	15.7	63.5
RP 2	17.7	14.1	68.2
RP 3	23.9	9.4	66.7
RP 4	19.3	13.0	67.7
RP 5	20.3	14.1	65.7
RP 6	20.7	11.5	67.8
RP 7	18.3	15.2	66.6
RP 8	20.7	11.3	68.1
RP 9	24.5	13.3	62.1
RP 10	19.7	14.2	66.1

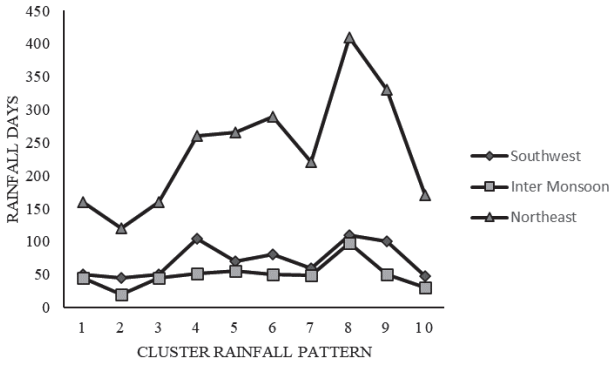

Fig. 6. Monsoons distribution for the ten rainfall patterns of torrential daily rainfall.

monsoon was recorded highest frequency of percentage distribution of torrential rainfall occurred in each rainfall patterns.

Fig. 6 illustrated that an accentuated maximum is observed in this torrential rainfall during northeast season (November to March). During this period, the winds over the east coast states of Peninsular Malaysia may reach 30 knots with strong surges of cold air from the north [26]. This is the most substantial monsoon for all RPs. Inter monsoon season loses its relative importance in this study as torrential events are rarely observed during April and October.

Conclusions

This study proposes a PCA based on Tukey's biweight correlation for identifying patterns of spatial heavy rainfall across Peninsular Malaysia. This proposed technique presents an alternative correlation matrix that addresses issues related to non-Gaussian distributed data, especially in light of skewed hydrological data. More substantial improvement was noted for partition of clusters by the proposed PCA method than the Pearson-based PCA, so as to prevent yielding imbalance and misleading clusters within space with high dimensionality. Besides, quality of the clustering results was determined based on three validity indices. The proposed PCA method was backed by simulation outcomes, which displayed more substantial improvement for partition of cluster than the Pearson-based PCA did to determine the pattern of special heavy rainfall across Peninsular Malaysia. Comparing the ten maps presented, each of the patterns tends to highlight distinct locations and their areas of affected by torrential rainfall do not overlap exceedingly. It is quite evident that, in general overview, Terengganu is the location most affected by torrential rainfall events. Six of the ten rainfall patterns are indicated that the torrential rainfall patterns during the Northeast monsoon experiencing the heaviest rain in the Eastern region of the Peninsula. RP 7 is located

in Kg. Jabi (Terengganu) received maximum day of torrential rainfall and it is interesting to note that this rainfall pattern is the most frequent one.

Acknowledgements

This study was produced under the Fundamental Research Grants Scheme 2019-0132-103-02 (FRGS/1/2019/STG06/UPSI/02/4) offered by the Malaysian Ministry of Education.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Worst floods in Kelantan, confirms NSC. Available online: http://www.themalaymailonline.com/ (accessed on 30 December 2019)
2. KOUMARE I. Temporal/Spatial Distribution of Rainfall and the Associated Circulation Anomalies over West Africa. Pakistan Journal of Meteorology, **10** (20), 1, **2014**.
3. XU T., CROKE B., HUTCHINSON M. Identification of Spatial and Temporal Patterns of Australian Daily Rainfall under a Changing Climate. 7th International Congress on Environmental Modelling and Software, 1, **2014**.
4. LEE H.S. General Rainfall Patterns in Indonesia and the Potential Impacts of Local Sea on Rainfall Intensity. Water, **7**, 1751, **2015**.
5. MERABTENE T., SIDDIQUE M., SHANABLEH A. Assessment of Seasonal and Annual Rainfall Trends and Variability in Sharjah City, UAE, Advances in Meteorology, 1, **2016**.
6. CRISTIANI E., VELDHUIS M-C.T., GIESEN N.V.D. Spatial and Temporal Variability and Their Effects on Hydrological Response in Urban Areas- A review. Hydrology Earth System Science, **21**, 3859, **2017**.
7. SUN Q., MIAO C., DUAN Q., ASHOURI H., SOROOSHIAN S., HSU K-L. A Review of Global Precipitation Data Sets: Data Sources, Estimation and Intercomparisons. Review of Geophysics, **56**, 79, **2018**.
8. HUBER P.J. Robust Statistics. Canada: John Wiley & Sons. **1981**.
9. ROUSSEEUW P., LEROY A. Robust Regression and Outlier Detection. New York, USA: John Wiley and Sons, Inc. **1987**.
10. SHAHARUDIN S.M., AHMAD N., ZAINUDDIN N.H., MOHAMED N.S. Identification of Rainfall Patterns on Hydrological Simulation using Robust Principal Component Analysis. Indonesian Journal of Electrical Engineering and Computer Science, **11** (3), 1162, **2018**.
11. PENARROCHA D., ESTRELA M.J., MILLAN M. Classification of Daily Rainfall Patterns in a Mediterranean Area with Extreme Intensity Levels: The Valencia Region. International Journal of Climatology, **22**, 677, **2002**.
12. ASSENT I. Clustering High Dimensional Data. Data Mining and Knowledge Discovery, **2** (4), 340, **2012**.
13. RUI XU., DONALD C., WUNSCH II. Clustering. Hoboken, New Jersey: John Wiley & Sons, Inc. **2008**.
14. NATHAN R., WEINMANN E. Australian Rainfall and Runoff Discussion Paper: Monte Carlo Simulation Techniques. Australia: Engineering House, **2013**.
15. SHAHARUDIN S.M., ISMAIL S., NOR S.M.C.M., AHMAD N. An Efficient Method to Improve the Clustering Performance Using Hybrid Robust Principal Component Analysis-Spectral Biclustering in Rainfall Patterns Identification. IAES International Journal of Artificial Intelligence (IJ-AI), **8** (3), 237, **2019**.
16. OWEN M. Tukey's Biweight Correlation and the Breakdown. Thesis, Pomona College. **2010**.
17. CHANG L. Robust Lasso Regression using Tukey's Biweight Criterion. Technometrics, **60** (1), 36, **2017**.
18. NEWARE S., MEHTA K., ZADGAONKAR A.S. Finger Knuckle Identification using Principal Component Analysis and Nearest Mean Classifier. International Journal of Computer Applications, **7** (9), **2013**.
19. HOAGLIN D.C., MOSTELLER F., TUKEY J.W.Understanding Robust and Exploratory Data Analysis. New York, USA: John Wiley and Sons, Inc. **2000**.
20. MONTFORT M.A.J.V., WITTER J.V. The Generalized Pareto Distribution Applied to Rainfall Depths. Hydrological Sciences Journal, **31** (2), 151, **1986**.
21. DHORE K.A., NATHAN K.K., KHARE D., CHAUBE, U.C. Weekly Rainfall Prediction using the Standard Package SMADA. Proceedings of the International Conference on Water and Environment (WE-2003). Bhopal, India, 32, **2003**.
22. CHO H-K., BOWMAN K.P., NORTH G.R. A Comparison of Gamma and Lognormal Distribution for Characterizing Satellite Rain Rates from the Tropical Rainfall Measuring Mission. Journal of Applied Meteorology, **43**, 1586, **2004**.
23. SHAHARUDIN S.M., AHMAD N., NOR S.M.C.M. A Modified Correlation in Principal Component Analysis for Torrential Rainfall Patterns Identification. IAES International Journal of Artificial Intelligence (IJ-AI), **9** (4), 655, **2020**.
24. SHAHARUDIN S.M., AHMAD N. Design and Simulation Systems. In: Ali MSM, Sahlan S, Wahid H, Yunus MAM, Subha NAM and Wahap AR. 752. Singapore: Springer; 216, **2017**.
25. JOLLIFFE I.T., CADIMA J. Principal Component Analysis: A review and Recent Developments. Philosophical Transactions, **374** (2065), 1, **2016**.
26. NIZAMANI Z., THANG K.C., HAIDER B., SHARIFF M. Wind Load Effects on High Rise Buildings in Peninsular Malaysia. IOP Conference Series: Earth and Environmental Science, **140**, 1, **2018**.